УДК 004.9:025.4
ББК 78.364.4

ИНДЕКСАЦИЯ ПОЛНОТЕКСТОВЫХ ДОКУМЕНТОВ В ЭЛЕКТРОННЫХ БИБЛИОТЕКАХ

Шаронова Н. В.,

Кочуева З. А.,

Борисова Н. В.

Национальный технический университет "Харьковский политехнический институт", Украина, г. Харьков

Анотація. *Розглядаються питання забезпечення доступу до даних в інформаційному просторі бібліотеки, а також питання автоматичного індексування повнотекстових документів для вирішення завдання інтелектуалізації пошуку інформації. Запропоновано алгоритм індексації як спосіб автоматичного індексування за допомогою ключових слів повнотекстових документів.*

Ключові слова: *автоматичне індексування, інформаційна технологія, ключові слова, електронні бібліотеки.*

Информация является одним из самых значимых ресурсов, оказывающих воздействие на развитие общества, его культуру, науку. С давних времен библиотека была центром научной и культурной жизни людей.

Важную роль в современных библиотеках начинают играть интернет-технологии, позволяющие читателям работать с каталогами, базами данных и другими удаленными от библиотеки информационными ресурсами.

С внедрением информационных технологий все большее значение приобретает развитие системы электронных библиотек.

Одной из основных задач является оцифровка фондов библиотек, как современный формат наращивания информационного контента для использования его в целях приобретения знаний. В рамках данной задачи

актуальной является проблема адекватного автоматического индексирования документов и извлечения из них сопутствующей информации

Индексирование текстов актуализируется так же при создании информационно-поисковых систем, использующих в качестве критериев поиска набор ключевых слов. Сложность этой задачи в том, как “правильно” определить этот набор. На сегодняшний день существует довольно много различных вариантов поиска текстов (или их фрагментов) по ключевым словам. Конечно, каждый из них имеет свои достоинства и недостатки.

Проблема, связанная с индексированием текстов состоит в том, что от ключевых слов (индексов) требуется соблюдение, как минимум, двух взаимоисключающих принципов: ключевые слова должны как можно точнее идентифицировать текст; ключевые слова должны как можно более точно отражать содержание (смысл) текста.

В общем случае эта проблема однозначно не разрешима, хотя и существуют достаточно эффективные системы поиска (например, поисковые системы в Интернет). Однако автоматическое индексирование и поиск ключевых слов в полнотекстовых документах необходимо проводить не только в Интернет, но и в современных библиотеках, которые нарастающими темпами накапливают неструктурированные текстовые ресурсы. Причем объем накопленной текстовой информации может быть таким затруднительным, что задача подготовки их полного библиографического описания становится крайне затруднительной. Очевидна необходимость применения специальных решений, которые позволят специалисту библиотеки автоматизировать процесс обработки полнотекстовых документов.

Разработанная система полнотекстового поиска учитывает морфологические особенности русского языка и реализует поиск информации по индексу, хранящемуся в реляционной базе данных. Ее основная функция – проводить индексацию текстовых документов, чтобы

затем можно было быстро и эффективно производить поиск необходимой информации по ключевым словам.

Процедура индексации включает следующие этапы:

1) перекодировка при необходимости документа в KOI8-R;

2) разбиение документа на отдельные слова с запоминанием позиции каждого слова в документе;

3) нахождение нормальных форм слов (если у слова есть несколько нормальных слов, то все они учитываются);

4) фильтрация стоп-слов (согласно содержимому таблицы stopwords), а также слов, короче 3 символов (в большинстве случаев короткие слова не несут смысловой нагрузки, и ими можно пренебречь при индексации);

5) подсчет частоты появления каждой нормальной формы в документе;

6) запись полученного конкорданса в БД.

Сама структура такого индекса должна обеспечить не только быстрый, но и релевантный поиск. Для повышения релевантности используется распространенный подход: при формировании терминологической словарной базы конкретного документа сохраняется не только сам термин, но и частота его вхождения в документ. Поэтому при выполнении поиска можно упорядочить его результаты по частоте вхождения искомого термина в документ.

Список использованных источников

1. Алисейко З. А. Автоматизированное индексирование полнотекстовых документов ключевыми словами / З. А. Алисейко, О. В. Канищева // Вестник Херсонского национального технического университета. – 2007. – № 4. – С. 269–272.

2. Кочуева З. А. Индексирование полнотекстовых документов для задачи интеллектуального поиска информации по ключевым словам / З. А. Кочуева, Н. В. Борисова // Східно-Європейський журнал передових технологій. – 2014. – № 1/2. – С. 4-8.

3. Хайрова Н. Ф. Автоматизированные информационные системы: задачи обработки информации / Н. Ф. Хайрова, Н. В. Шаронова – Х. : ХГУ «НУА», 2002. – 120 с.

Аннотация. *Рассматриваются вопросы обеспечения доступа к данным в информационном пространстве библиотеки, а также вопросы автоматического индексирования полнотекстовых документов для решения задачи интеллектуализации поиска информации. Предложен алгоритм индексации как способ автоматического индексирования при помощи ключевых слов полнотекстовых документов.*

Ключевые слова*: автоматическое индексирование, информационная технология, ключевые слова, электронные библиотеки.*

Annotation. *The problems of access to data in the information space of the library and the issue of automatic indexing of full-text documents to the task of finding information intellectualization. The algorithm is presented as an indexing method for automatic indexing keyword full-text documents.*

Key words*: automatic indexing, information technology, keywords, electronic libraries.*